VIA MEDICA

REVIEW ARTICLE

Piotr Potemski[1], Joanna Połowinczak-Przybyłek[1], Rafał Wójcik[2], Marcin Kaczor[3]

[1]The Department of Chemotherapy, Copernicus Memorial Multidisciplinary Centre for Oncology and Traumatology, Lodz; Chemotherapy Clinic, Medical University of Lodz, Poland
[2]Aestimo, Krakow, Poland
[3]Jagiellonian University Medical College, Krakow, Poland

Critical appraisal of clinical trials in oncology — part II

Address for correspondence:
Prof. dr hab. n. med. Piotr Potemski
Klinika Chemioterapii Nowotworów
Uniwersytet Medyczny w Łodzi;
WWCOiT im. M. Kopernika w Łodzi
e-mail: piotrpo@mp.pl

Oncology in Clinical Practice
2019, Vol. 15, No. 3, 158–166
DOI: 10.5603/OCP.2019.0013
Translation: dr n. med. Dariusz Stencel
Copyright © 2019 Via Medica
ISSN 2450–1654

ABSTRACT

The article is the second part of papers presenting informations useful for an independent analysis of the value of published results of clinical trials in oncology. Based on selected examples of clinical trials, a few attempts of critical appraisal of clinical trial assumptions, construction, and interpretation of their results are given. Several non-inferiority trials are discussed. The paper provides examples of publications in which post hoc analyses, grouping of variables, and multiple comparisons were made. Examples of research with a controversial selection of patients and a comparator, as well as studies whose clinical significance of the obtained results is questionable are presented. The aim of our work is to draw the reader's attention to selected essential elements of clinical trials and the way of presenting their results in order to facilitate practitioners in the independent evaluation of available publications and rational use of clinical trial results in everyday practice in the future.

Key words: oncology, clinical trials, critical appraisal, publication analysis, research methodology, interpretation of results

Oncol Clin Pract 2019; 15, 3: 158–166

Introduction

The first part of the publication presented general information helpful for independent analysis of the value of published results of clinical trials in oncology. Unfortunately, the description of the methodology is often presented in publications in a very short form, and more details can be found only after reading the protocol, which is not always available. In addition, the time the practitioner can devote to critically evaluate a new publication is usually very limited. All this means that it is quite difficult for the reader, who is a practitioner rather than a specialist in the field of clinical trial methodology, to systematically assess all the elements that make up the reliability of a given study, even after a very careful reading of the publication resulting from a clinical trial. This paper provides practical examples of interpretation of selected clinical trials. For obvious reasons, the analyses presented cannot be a comprehensive assessment of the results of these studies but are only an attempt to draw the reader's attention to selected, but in the authors' opinion very important, elements that may affect the interpretation of the published results and their impact on clinical practice.

Non-inferiority studies

Due to a different methodology than that utilised in commonly-used superiority studies, the non-inferiority design usually causes inconvenience for clinicians. It assesses whether an intervention is not inferior, in terms of clinical efficacy, than the current standard of treatment. The basic element subjected to critical evaluation during interpretation of this type of study is the assumed delta value, determining the acceptable difference in the clinical effectiveness of interventions being compared. It can be defined, for

example, by determining the magnitude of maintaining the clinical effectiveness of the current treatment standard, based on the results of a historical study comparing the current standard with symptomatic treatment (ASPECCT study) or determining the upper limit of the confidence interval based on the value of a clinically acceptable difference of effects adopted as part of the consensus (e.g. IDEA study). When interpreting the results of this kind of studies, it is worth paying attention to how large differences can be assumed that are still considered acceptable.

ASPECCT study

The ASPECCT study was a prospective, non-inferiority, phase III clinical trial planned to prove that panitumumab monotherapy in patients with metastatic colorectal cancer (CRC), who previously received chemotherapy, could result in at least half of the efficacy of cetuximab expressed by increased overall survival (OS) as compared to the best supportive care (BSC) demonstrated in a historic phase III study [1]. Such a defined delta value demonstrating non-inferiority seems to be a very safe assumption, which is easy to confirm in clinical trial. The study that was referred to in this assumption was the CO.17 study, the results of which, in a population of patients without *KRAS* mutation, were published in 2008 [2]. In this study the hazard ratio (HR) for death in the cetuximab group compared to only symptomatic treatment was 0.55 (95% CI 0.41–0.74), and median OS were 9.5 and 4.8 months, respectively. In total, 1010 patients were included in the ASPECCT study, and in the first scheduled analysis the assumption was proven, showing that panitumumab maintained from 81.9 to 129.5% of the effect of cetuximab on OS.

IDEA study

The IDEA (International Duration Evaluation of Adjuvant Therapy) study is a prospectively planned, pooled analysis of individual data of patients with colon cancer participating in six randomized trials, comparing the efficacy of three-month adjuvant oxaliplatin chemotherapy (FOLFOX-4 or modified FOLFOX-6 or CAPOX) with standard treatment lasting half a year [3]. The reason for planning such a study was the desire to reduce adjuvant therapy-related toxicity (mainly polyneuropathy) that may adversely affect the activity and quality of life of radically treated patients.

The primary endpoint was DFS (disease-free survival) in a modified intent-to-treat population (randomised patients who received at least one dose of chemotherapy), which was achieved by a total of 12,834 out of 13,025 randomised patients. The delta value was set as the upper limit of 95% CI HR_{DFS} of 1.12. Therefore, if 95% CI HR_{DFS} exceeded 1.12, the null hypothesis cannot be rejected, which would mean that the shorter duration treatment is worse than the assumed value than the standard treatment. According to the authors of the study, this delta value was estimated to translate into a predicted reduction in a DFS rate after three years by a maximum of 2.7 percentage points, and this value was considered acceptable by the researchers. As a reminder, in another study in CRC (MOSAIC), which established a value of FOLFOX adjuvant chemotherapy, patients with stage III disease had a DFS rate of 72.2% after three years, compared to 65.3% in those receiving fluorouracil with calcium folinate [4]. More important, however, is the effect of FOLFOX expressed in HR_{DFS}, which in the MOSAIC study was 0.76 (95% CI: 0.62–0.92), which means that HR for relapse decreased by 24%, and the "true" value of this reduction (i.e. the value transferred to the so-called general population) was between 8% and 38%. The delta value adopted in the IDEA study corresponds to maintaining about 60% of the effect in HR_{DFS} found for the comparison of FOLFOX to fluorouracil with calcium folinate in patients with stage III CRC participating in the MOSAIC study. In the IDEA study, according to the randomisation design in the primary studies used, interventions were compared only for the duration of chemotherapy, but not the type of chemotherapy.

In the modified intent-to-treat population HR_{DFS} for treatment lasting three months *vs*. six months amounted to 1.07 (95% CI 1.0–**1.15**), the assumed value of the statistically significant level of p-value for the hypothesis of non-inferiority three-month treatment was 0.11, and the p-value for the superiority hypothesis of six-month treatment was 0.045. This means that the primary endpoint was not met, and it was not proven (with the adopted delta value) that the shorter treatment is not inferior to the standard one. There are occasionally assessments of the results of this study based on numerical values of survival rates after three years — 74.6% and 75.5%, respectively. According to this assessment, the difference in a DFS rate (0.9 percentage points) is too small to be clinically relevant. This interpretation of IDEA research results shows a complete misunderstanding of statistical methodology and is entirely incorrect.

Moreover, pre-planned subgroup analyses were of an exploratory nature and cannot be interpreted in isolation from primary results of the study. It was found that probably the type of chemotherapy (FOL-

FOX or CAPOX) affects the effectiveness of three--month treatment (p for the interaction test was 0.006, and after the adjustment for multiple testing it was 0.02). In the group of 5071 patients receiving CAPOX HR_{DFS} amounted to 0.95 (95% CI: 0.85–**1.06**), which would confirm the assumptions of non-inferiority. Unfortunately, as previously mentioned, in the included studies no randomisation was made depending on the type of chemotherapy, or even randomisation was not stratified according to the type of chemotherapy. These factors mean that a result related only to the CAPOX scheme can be completely accidental, especially since it is difficult to find medical justification for such an observation.

Subgroup analyses and multiple comparisons not previously planned

As presented in the first part of the publication, randomisation is a very important element of a properly designed and conducted clinical study. It ensures, with a sufficiently large population, an even distribution of various, also unknown, confounding factors. The lack of randomisation or partial loss of its effect, e.g. as a result of post hoc analyses of previously unplanned subgroups of patients, means that compared groups may significantly differ in the distribution of other significant prognostic features.

IDEA study

In the IDEA study discussed above, subgroups were initially defined depending on the T (T1–3 and T4) and N (N1 and N2) feature, and none of the assumptions of non-inferiority of treatment lasting three months were shown in any of them. However, when analysing post hoc results, two categories of recurrence risk were created: low (T1–3N1) and high (T4 or N2). In the low-risk category (7471 patients) the assumption of non-inferiority regardless of the type of chemotherapy was confirmed at borderline (HR_{DFS} = 1.01, 95% CI: 0.90–**1.12**); similarly it was confirmed in the group of patients (N = 2,852) receiving CAPOX and classified as low risk (HR_{DFS} = 0.85, 95% CI: 0.71–**1.01**). In all other groups, i.e. high risk, regardless of the type of chemotherapy, or FOLFOX-treated low-risk patients, non-inferiority assumptions could not be demonstrated. It should be taken into account that the evaluation depending on these categories was not planned before, and it was performed only after analysing the obtained results. This means that in these subgroups the unknown additional factors may play an important role, and due to this the results of post hoc analyses cannot be considered as formal proofs used to infer real differences in the effectiveness of interventions. In the opinion of the authors of this work, the only potentially useful suggestion resulting from these analyses may be the possibility to shorten the duration of CAPOX chemotherapy to three months in patients with T1––3N1 CRC in the case of poor treatment tolerance, as an alternative to reducing the oxaliplatin dose or its withdrawal and continuing therapy with fluoropyrimidine alone.

ASPECCT study

Even better, the problem of multiple comparisons and random results considered "statistically significant" is illustrated in an article published in 2016, which presents the updated results of the ASPECCT study and, among others, post hoc analysis depending on previous treatment with bevacizumab [5]. It was found that in a group of 258 patients who were previously treated with bevacizumab OS was longer when they received panitumumab, not cetuximab. Medians OS were 11.3 and 9.8 months, respectively (HR = 0.75, 95% CI: 0.58–0.97). This observation has led to attempts to promote panitumumab rather than cetuximab as the drug of choice in patients previously exposed to bevacizumab. This raises the question of how to explain the advantage of panitumumab over cetuximab in individuals who have previously been treated with bevacizumab, when the biological mechanism of action of both drugs is very similar. This is a good example of misinterpretation of observation results, the nature of which is probably accidental and should not be the basis for a change in clinical practice. Obviously, in such situations, the statistically significant p-value should be lower than the usual one (< 0.05) because it must take into account unplanned hypothesis multiple testing used in post hoc analyses (Bonferroni correction).

Presenting the results of previously unplanned comparisons means that many similar ones were most probably carried out in other subgroups, but only some of them were selected because the more post hoc analyses are carried out, the more likely it is that the outcome of any of them will be completely randomly "statistically significant".

A study "showing" the importance of the astrological zodiac signs in medicine

Very instructive examples of the apparent demonstration of non-existent relationships are two works published by their authors just to show readers the dangers resulting from making multiple comparisons and grouping post hoc variables [6, 7].

In the first one, the relationships between astrological zodiac signs and the 223 most frequent reasons for hospitalisation of the inhabitants of Canada were evaluated [6]. A group of over 10 million people was randomly divided into two groups: a cohort in which possible relationships were tested and an independent validation cohort. Two zodiac signs were found to be associated with a higher risk of hospitalisation compared to the remaining 10. In the validation cohort, the relationship between the two signs and the individual causes of hospitalisation was examined, and it was found that people born under the sign of Leo were significantly more frequently (p = 0.0447) hospitalised due to gastrointestinal bleeding, and those born under the sign of Sagittarius significantly more often (p = 0.0123) due to humerus fracture, compared to people born under the other signs of the zodiac. Obviously, after introducing adjustment for multiple testing, these apparent relationships disappeared.

Another study investigated the relationship between the zodiac sign and prognosis after myeloablative chemotherapy and allogenic haematopoietic stem cell transplantation in patients with chronic myelogenous leukaemia [7]. The survival probabilities of at least five years were analysed in a group of 626 patients depending on their zodiac sign, and numerical but not statistically significant differences were found. However, when individuals born under the sign of Aries, Taurus, Gemini, Leo, Scorpio, or Capricorn (a total of 317 patients) were separated and compared with the remaining group (309 patients), the difference was statistically significant (five-year survival 58% *vs.* 48%; p = 0.007). Moreover, after a multivariate analysis that took into account the possible impact of other known prognostic factors, the results of treatment of patients born under one of the zodiac signs mentioned above were still significantly better than in the remaining patients (p = 0.005). The authors concluded that this is an example of "proving" a non-existent correlation, and the observed "significant" dependencies are the result of grouping post hoc variables in order to obtain the greatest and "statistically significant" difference.

COU-AA-302 study

An example of a proper interpretation of the possible impact of multiple testing in relation to survival outcomes is the adoption of another threshold of statistical significance for the results of pre-planned, stepwise analyses of randomised clinical trials. For example, in the phase III COU-AA-302 study, which compared abiraterone acetate in combination with prednisone to placebo-prednisone combination in patients with metastatic castration-resistant prostate cancer not previously treated with docetaxel, the two primary endpoints were: radiographic progression-free survival (rPFS) and OS. A typical p-value of 0.05 was therefore divided into both endpoints by default — the statistical significance of the difference in rPFS required that the p-value should be less than 0.01, and in the case of OS — less than 0.04 [8].

It was planned that the OS assessment would be conducted in several stages (interim analyses) — after the occurrence of 15%, 40%, and 55% of the number of deaths required for the final analysis, respectively, and final analysis after the occurrence of at least 773 deaths (1,088 patients were included in the study). Due to multiple testing of drug effects on OS (with data cut-offs at different time points), the correction of borderline p-values required to establish statistical significance found in these stepwise analyses of differences was applied in accordance with the procedure described by O'Brian and Fleming. The first publication contained the final result of the rPFS analysis, which found a statistically significant difference between abiraterone and placebo (HR = 0.55, 95% CI: 0.45–0.62, p < 0.001) and the result of interim OS analysis after 43% of the required 773 events. It was found that the HR_{OS} was 0.75 (95% CI: 0.61–0.93, p = 0.01). Although the p-value was less than the required 0.04, only the adjusted p-value of 0.001 or less was determined to indicate the statistical significance of the OS difference in this stepwise analysis. The result of the next published interim analysis carried out after 56% of deaths was also not statistically significant (HR = 0.79, 95% CI: 0.66–0.95, p = 0.0151, required adjusted p-value = 0.0035) [9]. Only the final OS analysis after 96% of the 773 deaths revealed the effect of the drug on OS prolongation (HR 0.81, 95% CI 0.70–0.93, p = 0.0033), i.e. it was allowed to meet the second co-primary endpoint [10].

Selection of patients and comparators

The correct patient selection, appropriate to carry out the planned intervention, is an indispensable element of a well-designed and conducted study, and also allows extrapolation of the outcomes to a population that will be treated in real-life clinical practice. The use of a proper comparator, which is a key element of a well-conducted clinical trial, implies the use of therapy that is consistent with current clinical practice and generally accepted recommendations and guidelines, including their continuous evolution, especially in a field developing as rapidly as oncology. In-

appropriate selection, contrary to commonly accepted recommendations, makes it impossible to accept study conclusions as reliable (external credibility). An example of a recently published trial with highly controversial patient selection is the CARMENA study.

CARMENA study

The aim of the non-inferiority phase III CARMENA (Cancer du Rein Metastatique Nephrectomie et Antiangiogéniques) study published in 2018 was to show that not performing nephrectomy in patients with metastatic renal cell carcinoma (mRCC) prior to sunitinib treatment is not inferior to such therapy with previous nephrectomy [11]. The primary endpoint was OS, and randomisation was stratified, among others, by prognostic categories. The results of the CARMENA study are fairly widely interpreted as being likely to change clinical practice, as it has been shown that not performing cytoreductive nephrectomy is non-inferior (HR_{OS} = 0.89; 95% CI: 0.71–1.10; non-inferiority criterion: upper limit of 95% CI not higher than 1.20).

However, while analysing the significance of the obtained result and its potential impact on clinical practice, the key limitation of this study should be remembered, which was the selection of patients. The study included patients meeting the criteria of intermediate or poor prognosis according to the MSKCC (Memorial Sloan Kettering Cancer Centre), and as many as 43% of patients participating in CARMENA study were assigned to the category of unfavourable prognosis. Until recently, in patients included in the category of unfavourable prognosis, the only drug for which phase III study showed a slight effect on OS prolongation was temsirolimus, a drug currently being reimbursed in such patients, also in Poland [12]. There are also no reliable data from a randomised study confirming the effect of sunitinib, used in CARMENA study, on OS in patients with unfavourable prognosis. In addition, nephrectomy is generally not performed or recommended in such patients, and in the aforementioned phase III study with temsirolimus, no beneficial effect of this procedure (sometimes performed a long time prior to randomisation) was observed on the efficacy of mTOR inhibitor. For these reasons, allowing inclusion in the study assessing the impact of abandoning nephrectomy the patients with poor prognosis category and treating them with sunitinib should be considered as not justified by existing medical knowledge. The use of sunitinib but not temsirolimus in some centres in patients with poor prognosis cannot be considered as practice in line with Evidence-Based Medicine.

The results of the CARMENA study were obtained in all patients enrolled, and those with an unfavourable prognosis had a significant influence on the final study results. In this group of patients neither nephrectomy nor sunitinib could affect the primary endpoint. With this assumption, it would not be difficult to prove non-inferiority of not performing nephrectomy in high-risk patients. An indication that seems to support this hypothesis may be the results for each prognostic group separately. In the group of unfavourable and intermediate prognosis, HR_{OS} were 0.86 (95% CI: 0.62–**1.17**) and 0.92 (95% CI: 0.68–**1.24**), respectively. As can be seen, only in the unfavourable prognosis group did the obtained result meet the accepted non-inferiority criterion (upper limit of 95% CI not greater than 1.20). Obviously, this cannot be considered as evidence but only a premise indicating the correctness of the given interpretation.

The statistical analysis carried out in the intent-to-treat (ITT) population assumes the evaluation of the results of all randomised patients, regardless of whether they received the assigned intervention or not. The interpretation of the result of the CARMENA study should also take into account the fact that 16 patients in the group of 226 included in the nephrectomy arm (7%) did not have it, and 38 of 224 subjects randomised to the arm with only systemic therapy (17%) underwent nephrectomy. This reduces the differences between the arms and facilitates the demonstration of non-inferiority in the ITT population. Finally, the original plan assumed the inclusion of 576 patients and the evaluation after 452 deaths, but as a result of the unsatisfactory recruitment rate the study was discontinued after including 450 patients, and the final results were published after the occurrence of 326 deaths.

CheckMate 214 study

The problem of selection of the comparator and target population was also encountered in the phase III CheckMate 214 study, in which the combination of nivolumab with ipilimumab in patients with mRCC was evaluated, and sunitinib was used as a comparator [13]. Whereas in the case of patients with a favourable or intermediate prognosis such a comparator does not raise any doubts; in the case of patients included in the category of unfavourable prognosis it is difficult — for reasons discussed earlier — to be considered as optimal. Such patients accounted for as much as 21% of the population in which the primary endpoints were assessed, i.e. objective response rate, PFS, and OS. The p-value of the statistical significance 0.05 was divided into: 0.001 (objective response rate), 0.009 (PFS), and

0.04 (OS). Patients were included in the study regardless of the prognostic category, but the assessment of primary endpoints was only planned for patients with an intermediate or unfavourable prognosis. Secondary endpoints included: objective response rate (ORR), PFS and OS in randomized patients (IT population), and the frequency of adverse events in patients who received treatment. An exploratory analysis was planned only in a group of 249 patients with a favourable prognosis (21% of the randomised population).

Regarding the primary endpoints, there were differences in ORR and OS in favour of immunotherapy, but not in PFS ($p = 0.03$). In the randomised population, no significant difference was found in any of the three secondary efficacy endpoints. This means that the inclusion of patients from the favourable prognosis group abolished the beneficial effect of immunotherapy with respect to response rate and OS. The results of an exploratory efficacy analysis in a favourable prognostic group are very worrying — there has been a reversal of the influence of immunotherapy and comparator to the detriment of experimental treatment. The response rate was 29% *vs.* 52% ($p < 0.001$), $HR_{PFS} = 2.18$ ($p < 0.001$), and $HR_{OS} = 1.45$ ($p = 0.27$, with only 37 deaths).

An obvious interpretation of the study results indicates that the benefit of immunotherapy is limited only to patients in the intermediate or poor prognosis category (with explicit reservation regarding external reliability due to the use of a suboptimal comparator in the latter group). However, it should be noted that only one of the six factors of unfavourable prognosis according to the International Metastatic Renal Carcinoma Database Consortium (IMDC) classification: Karnofsky performance status 70, time from diagnosis of cancer to randomisation shorter than one year, anaemia, corrected calcium concentration in serum above 10 mg/dl, neutrophilic granulocytosis, and thrombocythaemia, was associated with a benefit in immunotherapy. Inevitable doubts therefore arise as to whether this relationship is true for each of these mentioned factors that are so different, and whether the benefit of immunotherapy depends on the number of poor prognosis factors. Unfortunately, the publication of CheckMate 214 results does give any answers these question. Among 667 patients belonging to the intermediate-risk group, no analyses were performed that could clarify these doubts.

Another surprising choice was the use of only a combination of anti-CTLA4 and anti-PD1 antibody in the experimental arm, but not anti-PD1 monotherapy. This could be due to the desire to obtain the best direct effect (ORR was one of the primary endpoints). This does not undermine the value of the combination itself, but raises the question, however, of whether anti-PD1 monotherapy would not be equally effective and less toxic as well. This question can be considered as justified especially in the context of the final results of the previously launched CheckMate 025 study, published at the end of 2015, in which in previously treated patients nivolumab was used with good results. Such a doubt was raised by the European Medicines Agency (EMA) Committee for Medicinal Products for Human Use (CHMP) as justification for a surprising primary negative opinion regarding the registration of this combination in patients with kidney cancer.

SOLO3 study

It is also bewildering to select a comparator in the ongoing phase III SOLO3 study, in which olaparib is compared to single-agent chemotherapy in women with recurrent (at least two earlier lines of chemotherapy containing a platinum compound) still platinum-sensitive (progression later than six months after the end of the last chemotherapy) ovarian cancer with the presence of germline *BRCA* mutation [14]. The planned primary study endpoint is ORR. The comparator is the investigator's choice between paclitaxel, liposomal doxorubicin, topotecan, and gemcitabine.

As a reminder, a commonly accepted standard treatment for such patients is platinum-based chemotherapy and not monotherapy with any of the drugs mentioned above. In addition, the choice of the primary endpoint is also difficult to consider to be appropriate and clinically important. Considering these reservations, it is difficult to imagine that the result of this study could be useful in clinical practice.

20100007 study

If the use of placebo, or only BSC, as a comparator is common practice when there is no other therapeutic option with proven efficacy (usually the last treatment line), or the intervention tested and placebo as a comparator are added to the current standard (add-on), the use of placebo or only BSC in a situation where there are other therapies with previously demonstrated efficacy should always raise ethical concerns. The aim of this phase III study published in 2016 was to demonstrate that panitumumab affects OS prolongation compared to BSC in previously systemically treated patients with metastatic CRC without mutation in exon 2 of the *KRAS* gene [15]. There would be nothing surprising in the study design if not for the fact that the recruitment of patients was carried out between

November 2011 and July 2013, offering BSC as a comparator. This contradicted common knowledge about the efficacy of cetuximab in the treatment of patients with metastatic CRC. In the phase III study published in 2007 (the recruitment started in January 2004) and in which the value of panitumumab was for the first time evaluated in comparison to BSC, the authors highlighted that the project assumed from the beginning the possibility of changing the study arm in the control group after disease progression (cross-over) due to the "previously known activity of panitumumab and cetuximab" [16]. In addition, the design of this study with PFS as the primary endpoint and assumed cross-over meant that during the registration, the manufacturer was required to plan and conduct a non-inferiority study comparing panitumumab with cetuximab, because since 2007 the effect of cetuximab on OS prolongation has been known in comparison to BSC. Both antibodies, i.e. panitumumab and cetuximab, have been registered for the treatment of patients with chemoresistant metastatic CRC by both the US Food and Drug Administration and the EMA. In the European Union cetuximab was authorised in 2004 and panitumumab in 2007 (already taking into account the state of the *KRAS* gene). In 2008, the registration of cetuximab was modified, taking into account the state of the *KRAS* gene. In February 2010, recruitment to the previously discussed non-inferiority ASPECCT study was started, in which OS was the primary endpoint. How then, a few years after the first registration of both drugs, was it possible to conduct a clinical trial in which half of the patients included in the study received only BSC or a suboptimal procedure? Obviously, in the 20100007 study cross-over to panitumumab in the control group after disease progression was not assumed, and yet such an option was in the study, which was conducted during the period when there was no data confirming the effect of anti-EGFR drugs on OS.

Is every statistically significant difference also clinically relevant?

An element of critical evaluation of clinical trials in oncology should always be an answer to the question of whether the statistical significance obtained in a study is of practical significance and whether it is enough to change clinical practice. This issue raises a lot of controversy, because the assessment of how much of the benefit from a PFS or OS extension can be considered clinically relevant is extremely subjective. The discussion about this began, among other things, because there was a tendency to design commercial studies carried out on very large groups of patients, in which very small differences in efficacy could be demonstrated, but still achieving a level of statistical significance. Taking into account the fact that the primary endpoint of these studies was PFS, it was difficult to translate these results into clinical practice, especially considering the higher toxicity and significant cost of new drugs.

NCIC CTG PA.3 and VELOUR studies

The phase III NCIC CTG PA.3 study showed that the combination of gemcitabine and erlotinib in patients with advanced pancreatic cancer statistically significantly prolongs OS [17]. Although it was the first phase III study indicating advantage of combining gemcitabine with another drug, a fairly common perception of the clinical relevance of this study was far from enthusiastic, and this study became a classic example of a statistically significant but clinically meaningless result. The reason for this was primarily the low numerical difference in OS — HR had a value of 0.82 (95% CI: 0.69–0.99), the median OS 6.24 *vs.* 5.91 months, and 12-month survival rate was 23% *vs.* 17%. The large number of patients included in the study (569 people) meant that the absolute difference in the number of deaths between the arms of eight cases translated into statistical significance in the log-rank test. Especially underlined by the commentators was the difference in medians amounting to only 0.33 months. In addition, an increased frequency of some adverse events, e.g. diarrhoea and skin lesions, was observed in the experimental arm.

Assessment of the NCIC CTG PA.3 study value tested only from the perspective of difference in medians, although easy to communicate, is obviously somewhat simplified because it covers only one time point on survival curves. A better, although non-intuitive, measure is HR for death, in which an 18% reduction is already more clinically promising. For comparison, aflibercept, a drug currently being reimbursed in Poland in second-line treatment for patients with metastatic CRC added to FOLFIRI chemotherapy regimen in the phase III VELOR study, reduced HR for death by 18% (HR 0.817, 95% CI 0.713–0.937). The difference in medians was 1.44 months, and the probability of 24-month survival was 28% *vs.* 19% [18]. Demonstration that such a difference is statistically significant at the p level of 0.0032 was possible due to the inclusion of up to 1,226 patients in the study. Then, several reports were published to dispel doubts as to whether the difference in prognosis found in the VELOUR study is clinically relevant. One year after the original publication, extrapolating the obtained data

beyond the duration of the study using mathematical methods, the mean survival times of patients from both arms were estimated in the perspective of 15 years, stating that the difference between them is 4.7 months, which seems to have made an improved impression on the readers of that time than the difference in medians of 1.44 months [19]. Another attempt, the result of which was announced in print in 2014, consisted of making post hoc analyses and, as a result, extracting the subgroups referring to the "greater" benefit from experimental treatment [20]. It was found that patients in very good performance status (PS 0) with any number of distant metastases and patients in good condition (PS 1) having metastasis only in one location have a median OS greater by 3.1 months if they received aflibercept. Obviously, such analyses can generate research hypotheses, but certainly they do not allow the transfer of the results obtained in this way to the so-called general population. The value of post hoc analyses based on variable grouping has already been discussed in this article. It should be mentioned here that such actions are unfortunately used to obtain the greatest possible difference in the median of survival, which facilitates obtaining more favourable results of the cost-effectiveness analysis carried out as part of the process of reimbursement application. It is very likely that the subgroups extracted in this manner may not have any real predictive value.

NO16966 study

An example of a study that, at least some countries (e.g. in the USA), influenced the change of clinical practice despite seriously doubting the real value of the obtained results, was a phase III trial evaluating the benefit of adding bevacizumab to oxaliplatin-containing chemotherapy in the first line of treatment of patients with metastatic CRC [21]. The primary endpoint was PFS, and patients who received either FOLFOX-4 or XELOX chemotherapy with bevacizumab showed longer PFS (HR = 0.83, 95% CI: 0.72–0.95, median PFS: 9.4 *vs.* 8.0 months). The difference was statistically significant at p value = 0.0023, and it was possible to demonstrate it due to the sample size of 1,401 patients. Naturally, there was no OS benefit due to the use of antibody.

Doubts about the clinical significance of the results of some studies are the reason that ESMO (European Society for Medical Oncology) proposed the values of differences in individual endpoints depending, among others, on a prognosis that may be considered clinically significant [22]. The review of randomised clinical trials published between 2011 and 2015 regarding systemic treatment of patients with breast cancer, NSCLC, CRC, or pancreatic cancer included 277 studies [23]. In 138 of these studies, statistically significant differences between experimental therapy and the comparator were presented; however, after using the ESMO criteria of clinical significance, the results only 43 (31%) out of 138 studies were considered to be statistically significant.

Summary

In this analysis, the authors focused on selected issues, illustrating them with examples of specific clinical trials. Non-inferiority studies have been discussed because this type of clinical trial usually poses a lot of problems to readers, which is associated with a completely different methodology compared to studies that aim to demonstrate the superiority of one intervention over another. Examples of publications with post hoc analyses, grouping of variables, and multiple comparisons are given. Examples of clinical trials are presented, understanding and interpretation of which are impossible without paying attention to doubts about the characteristics of patients being included or the selection of a comparator. An extreme example of research with results that are difficult to transfer to clinical practice are those in which the control group is treated suboptimally, i.e. less effectively than is possible. Fortunately, there are not many of such studies, but more often there are clinical trials in which doubts relate to some of the patients included in them. Finally, examples of studies raising doubts about the so-called clinical relevance of the results obtained are given.

The authors hope that two publications prepared in cooperation of medical statisticians and oncologists will make easier for readers to interpret the available publications and thus rationally use the results of clinical trials in everyday practice.

References

1. Price TJ, Peeters M, Kim TW, et al. Panitumumab versus cetuximab in patients with chemotherapy-refractory wild-type KRAS exon 2 metastatic colorectal cancer (ASPECCT): a randomised, multicentre, open-label, non-inferiority phase 3 study. Lancet Oncol. 2014; 15(6): 569–579, doi: 10.1016/S1470-2045(14)70118-4, indexed in Pubmed: 24739896.
2. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med. 2008; 359(17): 1757–1765, doi: 10.1056/NEJMoa0804385, indexed in Pubmed: 18946061.
3. Grothey A, Sobrero AF, Shields AF, et al. Duration of adjuvant chemotherapy for stage III colon cancer. N Engl J Med. 2018; 378(13): 1177–1188, doi: 10.1056/NEJMoa1713709, indexed in Pubmed: 29590544.
4. André T, Boni C, Mounedji-Boudiaf L, et al. Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer. N Engl J Med.

2004; 350(23): 2343–2351, doi: 10.1056/NEJMoa032709, indexed in Pubmed: 15175436.
5. Price T, Kim TW, Li J, et al. Final results and outcomes by prior bevacizumab exposure, skin toxicity, and hypomagnesaemia from ASPECCT: randomized phase 3 non-inferiority study of panitumumab versus cetuximab in chemorefractory wild-type KRAS exon 2 metastatic colorectal cancer. Eur J Cancer. 2016; 68: 51–59, doi: 10.1016/j.ejca.2016.08.010, indexed in Pubmed: 27716478.
6. Austin PC, Mamdani MM, Juurlink DN, et al. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. J Clin Epidemiol. 2006; 59(9): 964–969, doi: 10.1016/j.jclinepi.2006.01.012, indexed in Pubmed: 16895820.
7. Szydlo RM, Gabriel I, Olavarria E, et al. Sign of the Zodiac as a predictor of survival for recipients of an allogeneic stem cell transplant for chronic myeloid leukaemia (CML): an artificial association. Transplant Proc. 2010; 42(8): 3312–3315, doi: 10.1016/j.transproceed.2010.07.036, indexed in Pubmed: 20970679.
8. Ryan CJ, Smith MR, de Bono JS, et al. COU-AA-302 Investigators. Abiraterone in metastatic prostate cancer without previous chemotherapy. N Engl J Med. 2013; 368(2): 138–148, doi: 10.1056/NEJMoa1209096, indexed in Pubmed: 23228172.
9. Rathkopf DE, Smith MR, de Bono JS, et al. Updated interim efficacy analysis and long-term safety of abiraterone acetate in metastatic castration-resistant prostate cancer patients without prior chemotherapy (COU-AA-302). Eur Urol. 2014; 66(5): 815–825, doi: 10.1016/j.eururo.2014.02.056, indexed in Pubmed: 24647231.
10. Ryan C, Smith M, Fizazi K, et al. Abiraterone acetate plus prednisone versus placebo plus prednisone in chemotherapy-naive men with metastatic castration-resistant prostate cancer (COU-AA-302): final overall survival analysis of a randomised, double-blind, placebo-controlled phase 3 study. Lancet Oncol. 2015; 16(2): 152–160, doi: 10.1016/s1470-2045(14)71205-7.
11. Méjean A, Ravaud A, Thezenas S, et al. Sunitinib Alone or after Nephrectomy in Metastatic Renal-Cell Carcinoma. N Engl J Med. 2018; 379(5): 417–427, doi: 10.1056/NEJMoa1803675, indexed in Pubmed: 29860937.
12. Hudes G, Carducci M, Tomczak P, et al. Global ARCC Trial. Temsirolimus, interferon alfa, or both for advanced renal-cell carcinoma. N Engl J Med. 2007; 356(22): 2271–2281, doi: 10.1056/NEJMoa066838, indexed in Pubmed: 17538086.
13. Motzer RJ, Tannir NM, McDermott DF, et al. CheckMate 214 Investigators. Nivolumab plus ipilimumab versus sunitinib in advanced renal-cell carcinoma. N Engl J Med. 2018; 378(14): 1277–1290, doi: 10.1056/NEJMoa1712126, indexed in Pubmed: 29562145.
14. https://clinicaltrials.gov/ct2/show/NCT02282020.
15. Kim TW, Elme A, Kusic Z, et al. A phase 3 trial evaluating panitumumab plus best supportive care vs best supportive care in chemorefractory wild-type KRAS or RAS metastatic colorectal cancer. Br J Cancer. 2016; 115(10): 1206–1214, doi: 10.1038/bjc.2016.309, indexed in Pubmed: 27736842.
16. Van Cutsem E, Peeters M, Siena S, et al. Open-label phase III trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. J Clin Oncol. 2007; 25(13): 1658–1664, doi: 10.1200/JCO.2006.08.1620, indexed in Pubmed: 17470858.
17. Moore MJ, Goldstein D, Hamm J, et al. National Cancer Institute of Canada Clinical Trials Group. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. J Clin Oncol. 2007; 25(15): 1960–1966, doi: 10.1200/JCO.2006.07.9525, indexed in Pubmed: 17452677.
18. Van Cutsem E, Tabernero J, Lakomy R, et al. Addition of aflibercept to fluorouracil, leucovorin, and irinotecan improves survival in a phase III randomized trial in patients with metastatic colorectal cancer previously treated with an oxaliplatin-based regimen. J Clin Oncol. 2012; 30(28): 3499–3506, doi: 10.1200/JCO.2012.42.8201, indexed in Pubmed: 22949147.
19. Joulain F, Proskorovsky I, Allegra C, et al. Mean overall survival gain with aflibercept plus FOLFIRI vs placebo plus FOLFIRI in patients with previously treated metastatic colorectal cancer. Br J Cancer. 2013; 109(7): 1735–1743, doi: 10.1038/bjc.2013.523, indexed in Pubmed: 24045663.
20. Chau I, Joulain F, Iqbal SU, et al. A VELOUR post hoc subset analysis: prognostic groups and treatment outcomes in patients with metastatic colorectal cancer treated with aflibercept and FOLFIRI. BMC Cancer. 2014; 14: 605, doi: 10.1186/1471-2407-14-605, indexed in Pubmed: 25142418.
21. Saltz LB, Clarke S, Díaz-Rubio E, et al. Bevacizumab in combination with oxaliplatin-based chemotherapy as first-line therapy in metastatic colorectal cancer: a randomized phase III study. J Clin Oncol. 2008; 26(12): 2013–2019, doi: 10.1200/JCO.2007.14.9930, indexed in Pubmed: 18421054.
22. Cherny NI, Sullivan R, Dafni U, et al. A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS). Ann Oncol. 2015; 26(8): 1547–1573, doi: 10.1093/annonc/mdv249, indexed in Pubmed: 26026162.
23. Del Paggio JC, Azariah B, Sullivan R, et al. Do Contemporary Randomized Controlled Trials Meet ESMO Thresholds for Meaningful Clinical Benefit? Ann Oncol. 2017; 28(1): 157–162, doi: 10.1093/annonc/mdw538, indexed in Pubmed: 27742650.